Análisis de conectividad: otra alternativa complementaria para la búsqueda y recuperación de información en la Wold Wide Web

Cristian Merlino Santesteban
Centro de Documentación. Facultad de Cs. Económicas y Sociales
Universidad Nacional de Mar del Plata
csantest@mdp.edu.ar

Palabras clave: análisis de enlaces, recuperación de información, topología de red, world wide web
Keywords: link analysis, information retrieval, network topology, world wide web

La World Wide Web (WWW, W3 o Web) es una red hipertextual de enorme complejidad que crece a un ritmo fenomenal. Esta inmensa estructura provee patrones de conectividad que pueden mejorar sustancialmente los métodos de búsqueda y ordenación por relevancia de los sistemas de recuperación.

Inspirado en el estudio de redes sociales y en el análisis de citaciones de la literatura científica, el uso de la estructura de enlaces o hipervínculos (*links*) ha emergido recientemente como un nuevo y promisorio acercamiento a la recuperación de información efectiva y eficiente.

La WWW puede ser vista estructuralmente como un grafo dirigido, donde cada página web es un nodo y cada enlace es un arco. La conectividad del digrafo, además de permitir la navegación, puede ayudar a caracterizar los documentos web.

El análisis de hipervínculos se basa en las siguientes suposiciones:

* Si un enlace de la página A enlaza a la página B, el autor de A está recomendando a B.

* Si la página A y B no estuvieran conectadas por un enlace la probabilidad de que ambas páginas tratasen el mismo o parecido tema sería mucho menor.

Algoritmos de ordenación por relevancia: dos casos exitosos

PageRank (Brin y Page, 1998)

El motor de búsqueda Google emplea un método de ordenación por relevancia independiente de la consulta del usuario denominado PageRank. El PageRank de una página web se calcula, a partir de un grafo generado *off-line*, ponderando a cada hipervínculo entrante con un peso proporcional a la autoridad de la página fuente.

HITS (Kleingberg, 1998)

El algoritmo HITS (*Hyperlink-Induced Topic Search*) propone una diferente noción de importancia de páginas web al considerar tanto los vínculos salientes como entrantes para identificar comunidades mutuamente relacionadas de páginas autoridad y *hub.*

Autoridad: nodo que recibe muchos enlaces. Recurso valioso.
Hub: nodo que vincula a muchas páginas web valiosas en un tópico.

Brin, Sergey y Page, Lawrence. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30, 107-117.
Kleinberg, Jon M. (1998). *Authoritative sources in hyperlinked enviroment*. Proceedings of ACM-SIAM Symposium on discrete algorithms, 668-677.